European Journal of Cardio-Thoracic Surgery 48 (2015) 180–193
doi:10.1093/ejcts/ezv168 Advance Access publication 12 May 2015

GUIDELINE

Cite this article as: Hickey GL, Dunning J, Seifert B, Sodeck G, Carr MJ, Burger HU *et al*. Statistical and data reporting guidelines for the *European Journal of Cardio-Thoracic Surgery* and the *Interactive CardioVascular and Thoracic Surgery*. Eur J Cardiothorac Surg 2015;48:180–93.

Statistical and data reporting guidelines for the *European Journal of Cardio-Thoracic Surgery* and the *Interactive CardioVascular and Thoracic Surgery*

Graeme L. Hickey[a,b,c,*], Joel Dunning[d], Burkhardt Seifert[e], Gottfried Sodeck[f], Matthew J. Carr[g], Hans Ulrich Burger[h] and Friedhelm Beyersdorf[i] on behalf of the *EJCTS* and *ICVTS* Editorial Committees

[a] Department of Epidemiology and Population Health, Institute of Infection and Global Health, University of Liverpool, The Farr Institute@HeRC, Liverpool, UK
[b] National Institute for Cardiovascular Outcomes Research (NICOR), University College London, London, UK
[c] Academic Surgery Unit, University of Manchester, Manchester Academic Health Science Centre, University Hospital of South Manchester, Manchester, UK
[d] Department of Cardiothoracic Surgery, James Cook University Hospital, Middlesbrough, UK
[e] Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zürich, Zurich, Switzerland
[f] Department of Emergency Medicine, Medical University Vienna, Vienna, Austria
[g] University of Manchester, Institute of Brain, Behaviour and Mental Health, Manchester, UK
[h] Hoffmann-La Roche AG, Basel, Switzerland
[i] Department of Cardiovascular Surgery, Freiburg University Heart Center, Freiburg, Germany

* Corresponding author. Department of Epidemiology and Population Health, University of Liverpool, Institute of Infection and Global Health, The Farr Institute@HeRC, Waterhouse Building (Block F), 1-5 Brownlow Street, Liverpool L69 3GL, UK. Tel: +44-151-7958306; e-mail: graemeleehickey@gmail.com (G. Hickey).

Received 27 March 2015; accepted 2 April 2015

Abstract

As part of the peer review process for the *European Journal of Cardio-Thoracic Surgery* (EJCTS) and the *Interactive CardioVascular and Thoracic Surgery* (ICVTS), a statistician reviews any manuscript that includes a statistical analysis. To facilitate authors considering submitting a manuscript and to make it clearer about the expectations of the statistical reviewers, we present up-to-date guidelines for authors on statistical and data reporting specifically in these journals. The number of statistical methods used in the cardiothoracic literature is vast, as are the ways in which data are presented. Therefore, we narrow the scope of these guidelines to cover the most common applications submitted to the *EJCTS* and *ICVTS*, focusing in particular on those that the statistical reviewers most frequently comment on.

Keywords: Guidelines • Statistics • Data • Reporting • Peer-review

INTRODUCTION

The *European Journal of Cardio-Thoracic Surgery* (*EJCTS*) and the *Interactive CardioVascular and Thoracic Surgery* (*ICVTS*) receive more than 3000 papers per year, many of which feature data and statistical analyses. The level of reporting varies from the simple presentation of data in tabular format, to the development of a clinical risk prediction model. Regardless of whether the manuscript was co-authored by a biostatistician or not, the majority of these papers are reviewed by a statistical consultant experienced in cardiothoracic and cardiovascular surgical data. This is part of the peer-review process that every research article undergoes in these journals [1].

Up until now the *EJCTS* and *ICVTS* have asked authors to follow guidelines for data reporting and nomenclature published in 1988 [2] and also the International Committee of Medical Journal Editors (ICMJE) Uniform Requirements for Manuscripts (http://www.icmje.org/icmje-recommendations.pdf; 29 January 2015, date last accessed). While these are excellent guidelines that authors should continue to use, we want to build on these for the specific *EJCTS* and *ICVTS* readership, with a view towards improving the quality of research published.

Here, we present new guidelines for authors on statistical and data reporting. These are built on our experience of reviewing manuscripts and from the guidelines of other biomedical journals [3], and as such represent only one view of what is considered important. It is emphasized that these are only guidelines and not strict rules. In quite a few places the guidelines are more 'what not to do' than 'what to do', which is in response to common errors. Moreover, this is not a guide on how to perform statistical analyses or choose appropriate methodology—for that one should consult an experienced statistician (preferably at the study design stage)—but rather on the presentation and minimum reporting required. There are a number of comprehensive textbooks on the topic of medical statistics that readers might consider for studies that require standard analyses [4, 5]. In places, however, we do direct readers to appropriate references where it is considered beneficial to authors, especially in circumstances where standard required methodology is frequently overlooked or underreported (e.g. model fit diagnostics).

The number of statistical methods used in the cardiothoracic literature is vast, as are the ways in which data are presented. Therefore, we narrow the scope of these guidelines to cover the most common applications submitted to the *EJCTS* and *ICVTS*, focusing in particular on those that the statistical reviewers most frequently comment on. This is evident from our bias towards commenting on Kaplan–Meier curves. In fact, by virtue of the scope and readership of the *EJCTS* and *ICVTS*, these guidelines are extremely narrow indeed.

Adoption of these guidelines should, in principle, lead to a clearer manuscript, thus allowing the reviewers to focus on what really matters—the science.

INTRODUCTION SECTION

Title

Titles should make clear the study design. Ostentatious or 'catchy' titles for manuscripts are to be avoided in reports with substantial analytical content. For example, 'A propensity-score matched analysis of 376 patients comparing surgery to no medical treatment for very mild aortic stenosis' is more informative than '*Much ado about nothing*: Should we operate on patients with aortic stenosis?'.

Aims and objectives

- Clearly state the objectives of the study. In particular, where appropriate, state any primary hypotheses that are to be tested.
- The authors should state clearly if the study hypotheses were prespecified or not.

MATERIALS AND METHODS SECTION

Study design

- Most studies conform to a standard design; for example, a randomized controlled trial or meta-analysis. Guidelines for standardizing, strengthening and increasing the transparency in the reporting of biomedical research in these areas are freely available and have been published. All authors should utilize these statements, which generally come with downloadable and printable checklists. If utilized, authors should declare which statement(s) they adhered to. Guidelines relevant to authors publishing in the *EJCTS* and *ICVTS* are available from:
  - *CONSORT* ['Consolidated standards of reporting trials'] statement: http://www.consort-statement.org.
  - *MOOSE* ['Meta-analysis of observational studies in epidemiology'] statement: reference [6].
  - *PRISMA* ['Preferred reporting items for systematic reviews and meta-analyses'] statement: http://prisma-statement.org/.
  - *STARD* ['Standards for the reporting of diagnostic accuracy studies'] statement: http://www.stard-statement.org/.
  - *STROBE* ['Strengthening the reporting of observational studies in epidemiology'] statement: http://www.strobe-statement.org/.
  - *TRIPOD* ['Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis'] statement: http://www.tripod-statement.org/.
  - *SQUIRE* ['Standards for quality improvement reporting excellence'] statement: http://www.squire-statement.org/.
  - *TREND* ['Transparent reporting of evaluations with nonrandomized designs'] statement: http://www.cdc.gov/trendstatement/.
  - *CHEERS* ['Consolidated health economic evaluation reporting standards'] statement: reference [7].
  - *SPIRIT* ['Standard protocol items: recommendations for interventional trials'] statement: http://www.spirit-statement.org/.

For other study designs not listed here, authors should consult the Equator Network (http://www.equator-network.org/library/) for specialist reporting guidelines where available.

- The *source of the subjects* should be detailed including, where appropriate, details of inclusion and exclusion criteria, and over what time period the data were collected.
- A *rationale for sample sizes* should be provided. Merely stating what software was used is insufficient. The statistical method used for estimation of effect sizes or assessment of statistical significance should be evaluated, including in the latter case whether a one- or two-sided test was used; if pilot data were available and used; what was the expected or clinically meaningful effect size sought; and what statistical precision of estimation, or power if significance testing is being used, was required. The statistical reviewer should be able to validate the calculation from the details provided.
- Authors should recognize that *small sample sizes could preclude certain statistical methods*. For example, fitting a multivariable logistic regression model to 25 subjects with 9 outcome events and 10 predictors is not sensible. Sample sizes and statistical methods should always be considered during the study design phase.
- Any study involving the *randomization of subjects* to different treatment arms should state the technique used, including whether simple, block, stratified or covariate-adaptive methods were employed [8]. Details on the randomization process such as blinding [9] and who was responsible for the randomization and allocation should also be noted [10, 11].

Outcomes

- The study outcomes should be clearly stated and defined. Frequently reported outcomes include:
  - *Operative mortality*: it should be stated whether this is (i) all-cause or procedure-specific mortality; (ii) the associated time point, for example in-hospital mortality; 30-day mortality; 90-day mortality; 30-day mortality including patients who died after 30 days but without discharge; within-intensive care unit mortality etc. The terms 'early mortality' or 'operative mortality' should not be used without the inclusion of a definition in the 'Materials and Methods' section. We emphasize that operative mortality is a binary outcome, and as such the appropriate statistical analyses are restricted to such data.
  - *Late mortality*: it should be stated whether this is (i) all-cause or procedure-specific mortality; (ii) what the time origin is (generally this will be the time of surgery or randomization). We emphasize that late mortality is generally accepted as being recorded as a time-to-event outcome, and as such the appropriate statistical analyses are restricted to such data, namely *survival analysis*.
  - *Postoperative complications*: each complication should be clearly defined. If a composite outcome is used, for example 'major' and 'minor' complications, then the individual components of each should be stated. Generally, postoperative complications are taken to be binary or count data in a fixed time window

(e.g. postoperative stay), but might also be evaluated as a time-to-event outcome.
  - *Time to reintervention*: it should be stated (i) what interventions are included (for example, if measuring the time to reintervention following surgical aortic valve replacement, would a transcathether aortic valve replacement be counted, or are surgical procedures counted only?); (ii) the time origin, which will typically be the time of surgery unless, for example, a landmark analysis is used [12].
  - *Continuous measurements*: it should be stated when and how the measurements were taken. For example, interleukin-6 might have been measured by taking a blood sample 1 h before bypass and 1 h after removal of the cross clamp.
  - *Count data*: data capturing an integer number of events (which might be binary). For example, the number of nodes resected. Care should be taken to distinguish whether the authors are counting subjects or counting a variable within a subject, as this will dictate the appropriate statistical methodology.
  - *Ordinal data*: for example, mitral valve regurgitation grade following valve repair or replacement surgery. The scale, definitions and time of measurement should be presented.
  - *Longitudinal data*: measurements taken repeatedly over time for each patient, either at fixed and regular time intervals or at different time points for each patient. For example, measuring the aortic gradient in outpatient clinics. Times of data collection should be clear and presented in a suitable format (tabular if regular periods, subject-specific time-series otherwise).
- Details should be provided on the *collection* of outcome data. For example, late mortality data might be collected using (a combination of) direct contact with patients, data linkage to a national or regional death register, contact with primary physicians, telephone or postal questionnaire surveys. Outcomes might be collected actively in the case of, say, a cohort study or by interview in, say, the case of a case–control study.
- For *valve-related outcomes* including structural valve deterioration and thromboembolic events, authors should consult reference [13].
- For *transcatheter aortic valve implantation (TAVI)-related outcomes*, authors should consult reference [14].
- *Composite outcomes* can increase statistical efficiency, particularly in studies where individual outcome components have a low incidence [15]. They also attract strong criticism [16]. Composite endpoints should be defined during the study design, and not after, to avoid any perception of manipulation. Additionally, complete data on individual component outcomes should be reported separately to facilitate interpretation.
- *Consistency in outcome terminology is important.* When reporting late events, there are often differing opinions on whether these should be referred to as 'late' or 'mid-term' events. This is subjective and application dependent; however, we stress that consistency in terminology is required to avoid confusion. Similarly, authors might report the terms 'operative mortality' and 'in-hospital mortality'. It is not always clear whether these are the same or different outcomes; hence one (appropriately defined) term should be used consistently throughout.

Study variables

- Details on the *data acquisition* methods should be provided. For example in case-series analyses one might interrogate a hospital unit computer records system, or manually enter data from obtained patient records. In a laboratory based study, details of the experiment should be provided with specific reference to the data measurement process.
- *Definitions of study variables*, such as patient characteristics and perioperative data should be provided where appropriate. For example, it is not clear whether 'recent myocardial infarction' means within 30 or 90 days, or whether 'smoker' includes only currently active smokers or those who recently quit.
- *When referring to the European System for Cardiac Operative Risk Evaluation (EuroSCORE)*, authors should explicitly state whether they are referring to the additive EuroSCORE, logistic EuroSCORE or EuroSCORE II [17–19]. Furthermore, one should avoid the term 'log EuroSCORE', as 'log' universally means 'logarithm', not 'logistic'. In this case, use 'logistic EuroSCORE' throughout.

Registration of clinical trials

- In accordance with the Clinical Trial Registration Statement from the ICMJE (http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/) all clinical trials published in the *EJCTS* and *ICVTS* must be registered in a public trials registry at or before the onset of participant enrolment. For any clinical trials commencing prior to 1 January 2008, retrospective registration will be accepted. For details, see the ICMJE website or reference [20].

STATISTICAL METHODS SECTION

It is advisable that a statistician be consulted to ensure that statistical methods are adequately described and applied and results are correctly interpreted. This is especially important in cases where non-standard methodology was applied.

Core requirements

- *All statistical methods used should be identified*. It is not acceptable to merely report that a particular software package was used to analyse the data.
- *Potential selection bias* in non-randomized studies where group comparisons are to be made should be elucidated.
- *Comparisons in non-randomized studies*, in particular observational studies, are difficult to make. Statistical methods should be used to adjust as much as possible for possible selection bias. Methods include multivariable modelling, matched pairs analysis and propensity score methodology. It is important to note that these techniques do not guarantee adequate correction for selection bias, but are generally more reliable than no adjustment at all.
- In *randomized trials it is unnecessary to test for baseline differences* between treatment groups unless there are concerns regarding the randomization process. In a randomized trial we know that baseline differences are by chance and therefore testing does not add value [21]. Instead, the authors should evaluate differences with regard to their ability to introduce bias in the analysis. Multivariable analysis models adjusting for such factors are appropriate tools for such an investigation.
- *Routinely used statistical methods*, for example Student's *t*-test and Fisher's exact test, do not require extensive details. However, in cases where misinterpretation is possible, authors should seek to provide additional details; for example stating an independent samples *t*-test was used rather than a paired-samples *t*-test. It should also be made clear where each method was used and for what purpose.

- *Avoid reporting that 'tests were used as appropriate'*. It is not always clear what is meant by this, let alone what is appropriate. Clarification on what tests were used should always be explicit. Especially, do not offer two separate tests for the same hypothesis without a clear description of which test is used in which situation.
- *Advanced methods* that are not regularly published in the *EJCTS* and *ICVTS* require explanation. Authors should consider including appropriate biostatistical references and/or an online methodological appendix. When a more advanced method has been used instead of a valid, more familiar approach, an explanation should be provided.
- Statistical methods, especially advanced methods, rely on *statistical assumptions*. For example, a particular statistical hypothesis test may rely on distributional assumptions such as normality, which if violated might require a preprocessing data transformation or application of a non-parametric test. Comparisons between groups might rely on the assumption of equal variances. Other commonly overlooked assumptions are described in these guidelines at the relevant sections.
- *Statistical software* used should be referenced (either as an in-line reference or citation) including the version number. Recommended references for the most commonly used software are:
  - *SPSS*: http://www-01.ibm.com/support/docview.wss?uid=swg21476197
  - *SAS*: http://www.sas.com/en_us/legal/editorial-guidelines.html
  - *R*: http://cran.r-project.org/doc/FAQ/R-FAQ.html#Citing-R
  - *GraphPad Prism*: http://www.graphpad.com/guides/prism/6/user-guide/index.htm?citing_graphpad_prism.htm
  - *Stata*: http://www.stata.com/support/faqs/resources/citing-software-documentation-faqs/
  - *JMP*: http://www.jmp.com/support/notes/35/282.html

  Any advanced 'add on' packages should also be referenced as appropriate, with version numbers where available.
- *Outlier data* should not automatically be eliminated from an analysis unless there are additional reasons to cast doubt on the credibility of the data. The influence of outliers on inferences should, however, be examined and consideration given to robust statistical methods.
- Some tests and methods, e.g. *t*-tests, depend theoretically on the *distributional assumption of normality*. However, in many cases these methods are quite robust to slight violations of this assumption. Use of a pretest for normality (e.g. the Kolmogorov–Smirnov test) as an automated tool to choose between a parametric (e.g. *t*-test) or non-parametric (e.g. Mann–Whitney *U*-test) test for comparison can be misleading. This is because, with small sample sizes, many tests of normality have low power for rejecting the assumption of normality, leading to false confidence. On the other hand, normality tests can reject the null hypothesis under very slight departures from normality when sample sizes are large.
- When reporting that a *correlation* will be calculated, note that there are multiple measures, and the reader should be explicitly told which is reported. Standard choices include Pearson's sample correlation *r* and Spearman's rank correlation.
- *Phrasing and terminology* should be used correctly.
  - The term *'multivariable' is preferred over 'multivariate'* when referring to regression modelling with a single outcome variable and multiple explanatory variables. 'Multivariate' should be reserved for modelling more than one outcome variable [22]. Similarly, the term 'univariable' is preferred over 'univariate'. The term 'bivariate' should only be used in the special case of two outcomes (a special case of multivariate analysis).
  - The term *'non-parametric'* should refer to a statistical method, not data themselves. For example, the statement that 'the Mann–Whitney *U*-test was used when data were non-parametric' is incorrect. The Mann–Whitney *U*-test is a non-parametric test.
  - The terms *'incidence' and 'prevalence'* are routinely confused. *Incidence* is defined as the number of new cases of a characteristic in a population over some fixed time; e.g. number of new rheumatic fever cases over a year for a single hospital. *Prevalence* is defined as the proportion of a population that has the characteristic; e.g. the proportion of patients undergoing mitral valve repair (the 'population') with diabetes mellitus (the 'characteristic').
  - The term *'correlation'* is often reserved in statistics for analyses reporting correlation coefficients. Authors should attempt to avoid using it in cases where the terms 'association' or 'relationship' might be more appropriate, for example when commenting on an odds ratio.
  - When calculating survival using the Kaplan–Meier estimator, the phrase *'actuarial survival'* should not be used; simply write 'survival' [23]. If life-table methods are used, then 'actuarial survival' is an appropriate term.
  - The term *'significance'* should not be used colloquially, and should be reserved only for reference to *statistical* significance or *clinical* significance (sometimes referred to as clinical relevance or clinical importance), being explicit about which.
  - *Quantiles* (e.g. terciles, quartiles, quintiles, deciles, percentiles etc.) *define the cut-points, not the groups*. For example, if dividing BMI data into five groups using the quintiles, the groups would not be called quintiles, they would be called fifths [24].
- *Define all statistical acronyms and notation* at first use, including standard deviation (SD), standard error (SE), standard errors of the mean (SEM), confidence interval (CI), *r*, *R*, rho, beta.

Non-independent data

- *Paired data* arise in situations such as a patient having a measurement recorded twice, such as before and after intervention (e.g. patient-reported pain score); or from a matched case–control study. It is important to use an appropriate test that accounts for the paired nature of the data, and to clearly note this in the manuscript. For example, one might consider using a Wilcoxon signed-rank test or paired *t*-test in the case of continuous variables.
- *Repeated measures data* typically arise in serially recorded data at different time points. For example, interleukin-10 measurements from subjects might be recorded on postoperative days 1, 3, 5 and 7 for a series of patients who underwent bypass surgery either off-pump or on-pump. Ignoring the dependency, or correlation, between responses within individual subjects can reduce power and precision in some cases. Standard tests for analysing repeated measures data include repeated measures ANOVA and Friedman's test. In the case of the former, authors should describe clearly (i) the underlying model and whether any interactions with time were analysed; (ii) how any assumptions, e.g. sphericity, were evaluated. For further information, including how to present the results of a repeated measures ANOVA, consult reference [25]. More flexible statistical methodology that can also be used falls under the umbrella of longitudinal data analysis [26].
- When describing analysis methods, it should be clear what the *unit of analysis* is. For example, in thoracic research it is common

for measurements to be taken on several lymph nodes from the same patient. If the unit of interest is the patient, then the measurements might be considered clustered within the patient. That is, measurements taken from different lymph nodes in the same patient would likely be more similar than those taken from different patients. This is another type of repeated data. It is important to use methods appropriate for this clustering. For example use of random-effects models [27] or generalized estimating equations [26].

- Studies that measure a continuous variable at some baseline followed by a measure at some point afterwards, e.g. following surgery or 1-year following treatment, might calculate *percentage change scores* of the type: $100 \times (f - b)/b$, where f is the follow-up measure and b is the baseline measure. These percentage change scores are then often statistically compared between groups, e.g. open-surgery repair vs endovascular repair. In many situations, however, this is not generally recommended [28]. Analysis of covariance (ANCOVA) is the preferred approach for comparison, owing to its superior statistical power [29].

Hypothesis tests and multiple comparisons

- It should be explicitly stated which *hypotheses were prespecified* and which were data-driven. Limiting testing to those hypotheses that were identified *a priori* is preferred. *Post hoc* analyses have their own value, but the interpretation of such analyses is different and more challenging.
- In analyses where *post hoc analysis* is to be performed, the authors should make clear that generally such analyses are of only hypothesis-generating value. The authors could apply suitable statistical techniques that account for multiple testing. However, while this could account for some of the multiplicity, it is still unlikely to remove the exploratory nature of such analyses. For groups with a natural ordering (e.g. different dosages of an anticoagulant), data should be analysed using methods that test for trend [30].
- Merely stating that 'multiple comparisons were accounted for' or 'results have been adjusted for multiplicity' is insufficient information. The authors should *name the method used to adjust for multiplicity.*
- The *Mann–Whitney U-test* (or Wilcoxon-Mann–Whitney test) for comparing two continuous or ordinal variables is not a test for comparing the difference in medians except in a particular special case [31]. Therefore, authors should avoid making this claim.
- Authors routinely report that categorical data were compared using either *Pearson's χ^2 test or Fisher's exact test*, as appropriate. The statement 'as appropriate' should be spelled out precisely. Often it is an implicit reference to whether or not the expected cell frequencies are <5. There is now evidence to suggest that this rule of thumb is outdated. Current recommendations are to use Pearson's χ^2 test with an $(N - 1)/N$ correction factor (where N is the total sample size being included in the test) applied to the test statistic in situations where the expected cell frequencies are >1 [32]. Some statistical software packages allow (or even automate) the Yates' continuity correction factor. Such corrections should be reported if used.

Missing data

- *Don't ignore the issue!* Many studies will encounter missing data to some degree, which without first handling generally precludes statistical analysis, e.g. regression. It is also one of the main limitations to standard analyses, for example longitudinal analysis. A description of methods used should be given.
- *Missing data needs to be carefully described* and discussed, as they may limit the conclusions that can be made. At least in the presence of a considerable amount of missing data, several different methods or sensitivity analyses should be applied for the analysis in order to see if conclusions are robust against missing data.
- *Not including important variables* in analyses due to missing data is not appropriate.
- *Imputation of data by the mean or median* of observed values (in the case of continuous or ordinal variables with missing data), or by the mode of observed values (in the case of categorical variables) results in overconfidence (*P*-values are too small; CIs are shrunk). This is because it treats the imputed data as if they were actually observed measurements.
- *Case-complete and missing-indicator methods* are generally inferior to the multiple imputation method [33, 34].
- *Multiple imputation* is generally a better method than simple imputation rules. Using multiple imputation methodology requires authors to propagate the uncertainty through into the calculation of effect size SEs. Many statistical software packages now incorporate analytical routines that can automate this process [35–37]. However, multiple imputation methods, like all methods, have limitations.

Regression analyses (general overview)

The following provides information on how regression analyses should be specified and performed. This is specifically important when the objective of the study is linked to a regression analysis like a prognostic factor analysis or a prediction model. Otherwise, the amount of information provided could be more limited. For more sophisticated models the underlying methodology could be considerably more complex, and should therefore be understood when such models are applied and the results interpreted. In such cases it is strongly advisable to consult a statistician first.

- *Terminology* used in the context of regression analysis varies. In general, one is interested in modelling the relationship between an outcome variable (also often referred to as the dependent variable or response variable) and one or more explanatory variables (also often referred to as independent variables, predictors or covariates).
- In many studies, especially observational studies, the *outcome of interest will be associated with many other variables* (e.g. patient demographics and comorbidities) in addition to the primary variables of interest (e.g. treatment). These covariates should be included in the regression models where sample sizes permit. All variables included in a model, including any higher-order terms, should be listed.
- *Model development strategies* should be clearly described in such a way that an independent analyst could validate the reported results independently if they had access to the original data. In particular:
  - Was prescreening of variables for inclusion decided on the basis of univariable hypothesis tests or on clinical reasoning?
  - Was a stepwise regression model algorithm used? Or were all *a priori* identified variables included?
  - Were any variables 'forced' into the model, even where a model development algorithm was used?

- *If univariable testing is used as a prescreening method* for variable inclusion in a multivariable model, then in general a threshold of $P < 0.05$ is too strict and will introduce selection bias. Typically a higher threshold should be used, for example $P < 0.25$.
- *Stepwise regression* is a broad term that captures forward, backward and bidirectional regression algorithms [38]. It should be stated which approach was used and what the inclusion and/or exclusion criteria were for the algorithms, as these vary by statistical software packages. Merely stating what software was used is not sufficient.
- *Stepwise regression has a number of major limitations*, despite its appeal, that authors should be aware of before using such methods [39], especially in the context of small datasets [40]. For example (examples taken from reference [41]):
  - The selection is typically unstable, sensitive to small perturbations in the data. Addition or deletion of a small number of observations can change our chosen model markedly.
  - SEs of regression coefficients are negatively biased, CIs are too narrow, *P*-values are too small and R^2 or analogous measures are inflated.
  - Regression coefficients are positively biased in absolute value.
  - It has severe problems in the presence of collinearity.
- *Continuous explanatory variables often do not need to be categorized*. Categorizing a continuous predictor into intervals (including two intervals, known as dichotomization) can lead to a number of statistical inferential problems including bias and loss of power [42]. Sometimes there might be a reason to categorize a variable based on either statistical evidence or for clinical reasons. In such a situation, this should be reported, and suitable efforts for defining a good cut-point for dichotomization should be applied. However, in general predetermined cut-points are preferred.
- The *functional form* of explanatory variables should be correctly specified. For example, in a logistic regression model relating in-hospital mortality to body mass index (BMI), linearity is unlikely to hold between the log-odds of mortality and BMI because patients with either extremely low BMI (very underweight) or extremely high BMI (morbidly obese) will be expected to have a higher risk of mortality. One might model the hypothesized U-shape in this particular example by including a quadratic term.

  Assessing functional form varies by regression model; however, standard methods include:
  - Plotting explanatory variables against *residuals* for ordinary linear regression [43].
  - Plotting *smoothed Martingale residuals* (with and without the explanatory variable in the model) against the explanatory variable for Cox proportional hazards regression [44].
  - Use of *splines* [39] *or fractional polynomial transformations* [45].
  - *Replacing the continuous variable by a percentile-categorized variable*, and plotting midpoints against fitted model coefficients. Note that the categorical variable method is a crude assessment, and should not automatically be used in place of continuously modelled variables [46].
- *Multicollinearity* is an issue that can affect any regression model. In general, it can lead to an inflation in type II errors (increased SEs for estimated coefficients), which makes identification of predictors difficult, as well as misleading coefficients. There are numerous techniques available for the identification of multicollinearity (e.g. variance inflation factors), which should be considered [47].
- *Interaction terms* allow for two variables to have non-additive effects in a regression model. For example, a treatment effect might be different in smokers and non-smokers. Use of interaction terms is similar to fitting regression models to different subsets of the data (e.g. in smokers and non-smokers). Interaction terms—just as the case for subgroup analyses—should be specified in advance; *post hoc* searching can lead to spurious significant effects. It should be noted that models including interaction terms are generally more difficult to interpret, thus requiring authors to carefully present and interpret their results [48, 49].
- If the study objective is to test a specific hypothesis, then an *a priori* sample size calculation should made in most circumstances. Formulae for sample size calculations are readily available for standard regression models including linear and logistic regression [50] and Cox proportional hazards regression [51].

Linear regression

- Ordinary linear regression depends on a number of *assumptions that should be assessed*, which include (in addition to the others mentioned above):
  - *Linearity*: see the comment on 'functional form' above.
  - *Homogeneity*: also referred to as homoscedasticity; the variance of the errors should be independent of the explanatory variables.
  - *Independence of errors*: errors should be independent of one another. This is particularly important if samples contained repeated measurements on subjects, or if there is an obvious clustering (e.g. different hospitals).
  - *Normality*: if one is interested in making inferences about the model, particularly with small sample sizes, then it is often assumed that the errors should be normally distributed.
  - *Diagnostics tools* for the evaluation of these assumptions generally involve inspection of the residuals [43].
- *If diagnostic analyses suggest that a revision to the model is required*, e.g. a transformation to the outcome variable to stabilize the variance, then descriptions of the methods should be made available and results provided.
- The *coefficient of determination* (R^2) should be reported, stating whether it is the standard or adjusted value.

Logistic regression

- *It is not always sufficient to consider the sample size* (i.e. number of subjects) *alone* when considering a study design that will involve a multivariable logistic regression model. The number of events per variable (EPV) ratio is also important. Although the rule of thumb has been the requirement of 10 EPV, this is not a strict necessity and in general is application dependent [52, 53].
- There are a number of *goodness-of-fit tests and model diagnostics* for evaluating a logistic regression model fit [46, 54]. A brief description of what methods were used should be provided (with the results presented in the 'Results' section).

Cox proportional hazards regression and survival analysis

- The *proportional hazards assumption* is fundamental to the Cox model, and any (gross) violation of it can potentially result in misleading inferences [55]. All Cox proportional hazard regression models are therefore required to include details and results of how this assumption was evaluated. Standard approaches include:
  - smoothed scaled Schoenfeld residual plots [56];
  - complementary log–log Kaplan–Meier plots [57];

- the Grambsch–Therneau test [56];
- including time-dependent coefficients into the regression model [58].

- Where required, *models should be adjusted to compensate for any violation of the proportional hazards assumption*. Standard approaches include the stratified Cox proportional hazards regression model and time-dependent variables/coefficients, or use of a different model [59].
- Similarly to logistic regression, *the sample size should be considered in relation to the total number of events* (i.e. the number of subjects that reached the endpoint), not just the total number of subjects or the total follow-up time [13].
- The Cox proportional hazards regression model is ubiquitous in the biomedical literature. However, authors should consider the availability of other time-to-event regression models depending on their specific study, as theoretical arguments might imply more suitable models. For example, modelling the time to structural valve deterioration of implanted bio-prostheses can exploit the theoretical knowledge that the hazard will increase with time. *Parametric survival models* such as Weibull regression or other accelerated failure time models can be used, and offer potentially greater power [57].
- *Do not include fixed predictors in the model that are only observed after the time origin*. For example, if modelling the time to death and the time origin is the commencement of surgery, then do not include postoperative complications (e.g. stroke) or post-baseline measurements as a (static) predictor. It is not possible to 'reach' forward in time! A classic example involves heart transplantation. In the Stanford Heart Transplant Study, patients who were eligible for a transplant were followed until death or censorship [60]. The objective of the study was to assess whether patients who received a transplant live longer than those not receiving a transplant. Comparing the survival distributions between those who received a transplant at some point and those who did not is, however, not appropriate. Patients who received a transplant must have lived until the point of transplantation, and so must have contributed survival time in the non-transplant group [61]. In short, transplantation status is not known at the time of eligibility; therefore, we should not condition on this variable it in advance. In such situations one might consider:
  - including a time-dependent covariate [58];
  - using a landmark analysis [12];
  - not including measurements that occurred after the time origin.

  However, this would be application dependent, and would be guided by the study objectives.
- *If multiple measurements are recorded for one or more predictors*, then including all data using time-varying covariates can lead to more informative inferences. For example, in a follow-up study of patients undergoing aortic valve replacement, patients might return to clinic regularly to have valve gradient measurements taken. These gradient measures can be included in a model as a time-dependent covariate. Time-varying covariates can also be used in cross-over trials to model change in treatment [61].
- For modelling time to non-terminal events, for example time to reintervention, patients who die might be considered a *competing risk*. Similarly, when modelling the time to a cause-specific mortality, other causes of death can be considered as competing risks. Ignoring competing risks can lead to biased estimates. Standard methodology for overcoming this issue include summarizing the survival data using the cumulative incidence function [62].
- The *area under the receiver operating characteristic curve* is sometimes reported in the context of survival analysis. Ignoring the time-to-event nature (including the censoring) of the data and treating it as a purely binary outcome should be avoided. There are extensions of discriminatory ability measures that can be used, e.g. Harrell's *C*-statistic [63], and generalizations, e.g. time-dependent ROC curves [64].
- It is often of interest to *compare the survival of a sample of subjects to a demographically matched population*, typically obtained from national actuarial tables. The standard log-rank test is not appropriate for this comparison; however, there are a number of proposed one-sample survival tests [65].
- *The log-rank test does not compare survival at fixed time points*; it compares the distribution of survival times between two (or more) groups. Statements such as 'the survival at 5 years was 65% in Group A and 75% in Group B ($P = 0.012$)' are therefore unclear and misleading when the *P*-value is calculated from the log-rank test. Tests that calculate the difference in survival at fixed time points are available [57]; however, time points should be limited and defined at the study design stage.

Risk model development and validation

- There are a number of statistical challenges when *developing a risk prediction model*, which have been extensively described in the literature. Authors who are unfamiliar with the issues are encouraged to consult a guide on model development prior to any analysis [63, 66].
- All *developed risk prediction models should be validated*, which can be internal or external. Validation by arbitrary data splitting (e.g. 60%/40% training/validation split) should be avoided in preference to bootstrap resampling methods [39, 67, 68]. At the most intuitive level, splitting data are wasteful, especially when sample sizes are small to moderate.
- If an *additive risk score* is developed, it should be explicit on how the scores were derived from the statistical model [69].
- The *external validation* of risk prediction models should describe both the model calibration and discrimination [70, 71].
- When *assessing the model calibration*, there is a growing need to reconsider the suitability of the Hosmer–Lemeshow test, especially with large sample sizes [72–74]. Instead, consider complementing it with a range of other calibration tests and diagnostics, in particular with the support of calibration plots with overlaid smoothed calibration curves [70, 75]. For testing whether the calibration curve is 'ideal' consider either a recalibration regression test [67, 76] or Spiegelhalter's *Z*-test [77].
- *External validation of a risk prediction model requires a substantial effective sample size*. This should be incorporated into the design stage of validation studies [78].
- *Assessing a risk prediction model 'off label'* should be explicitly stated, and conclusions should not mislead readers that the performance was for the intended outcome. For example, the EuroSCORE II was developed to predict in-hospital mortality. 'Validating' the model in a sample of patients with the outcome of 30-day mortality would be 'off label', and so caution should be taken.

Propensity scores and matching

- The use of statistical hypothesis tests for *evaluating balance* before and after matching is widely criticized [79]. The simplest accepted approach is to calculate the standardized bias [also known as the

standardized difference (in means)] for each variable:

$$d = 100 \times (\bar{x}_1 - \bar{x}_2)/\sqrt{(s_1^2 + s_2^2)/2}$$

where $\bar{x}_i$ and s_i^2 denote the sample mean and variance of the data for group $i = 1, 2$. There are a number of other approaches that can be considered also [80, 81].

- Variations of the *standardized bias for other propensity score methods* including regression adjustment [82] and inverse probability treatment weights are available [83].
- *Evidence of imbalance* ($|d| > 10\%$; significant *P*-values) in baseline variables between propensity score-matched groups should not be overlooked [84]. Propensity score model development is an iterative process. Authors should consider including polynomial terms; interactions; or including additional confounding variables [85].
- At the time of writing, no dedicated statement on the reporting of propensity score-matching methods are known to the *EJCTS*. However, the systematic reviews by Austin contain a useful number of *recommendations on the minimum reporting requirements* [79]. These include describing:
  - the propensity score model development, including the prior selection of variables for inclusion in the model;
  - the matching algorithm used, along with details of any tuning attributes (e.g. caliper widths). This should also report whether 1:*M* matching was employed and whether matching was with or without replacement;
  - methods used for assessing balance;
  - statistical methods for estimation of treatment effects.
- There is *debate about whether treatment effects should be estimated using methods for matched data* or unmatched data. Austin vociferously argues that the correct statistical methods for estimation of treatment effects are those that account for the matching [79]. However, this has been rebutted by other experts [86, 87]. Frustratingly, there is no obvious right or wrong answer to this. Notwithstanding this lack of consensus, authors should nonetheless avoid interchanging the application of statistical methodology appropriate for matched and unmatched data when comparing different outcomes.
- *Goodness-of-fit tests* (e.g. the Hosmer–Lemeshow test) and the *C*-statistic are not generally informative about the adequacy of the propensity score model specification [81].
- Propensity score methodology is predominantly focused on the case of two treatment arms. There is a growing body of research into performing *propensity score matching for three or more groups*; however, they are foundationally and technically more challenging [88, 89]. An exceptionally well-detailed description of methods involving such sophisticated methodology would be required for review.
- *Including the propensity score in a multivariable regression model* with a treatment indicator variable is a common technique for estimation of treatment effects [90]. However, it does not preclude one from ensuring that regression model assumptions are satisfied. This includes ensuring that the outcome model is adequately specified in the same way one would any other multivariable model. If any violation of assumptions are detected, then quadratic terms, logit transformations or regression splines might be used to model the score correctly [84].
- When applying propensity score matching, *the unmatched subjects* (which can be in both treatment arms) should ideally also be profiled [91].
- *Matching* can be done outside of the propensity score paradigm, for example in matched case–control studies. The process by which matching was performed should be described, including: (i) whether 1-to-1 matching (or otherwise) was performed; (ii) what variables were used to match on, and if exact matching was used (e.g. how was matching handled for continuous variables?); (iii) in the cases of >1 match, what rules were applied?
- *When subjects are matched exactly on a set of variables*, then statistical tests that are appropriate to matched data are most efficient [92].
- Propensity score-matching methodology might be precluded in the cases of very small numbers of subjects in the treatment group.

Diagnostic studies

- *Report the known diagnostic sensitivity and specificity* of any test used with reference to the information source.
- *Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)* are all measures of the diagnostic ability of a test. The PPV and NPV are not intrinsic to the test—it depends also on the prevalence [93, 94].
- *Using specificity to compare two tests can mask important differences*. For example, specificities of 94 and 97% look high; however, the false-positive rates are 6 and 3%, respectively—the former is twice as large as the latter. Reporting the 2 × 2 classification table is advisable.

Receiver operating characteristic curves

- When *comparing the area under two (or more) ROC curves* calculated from the same data, it is necessary to account for correlation [95, 96].
- *Calculating cut-points from ROC curve analysis* requires reporting what method was used. There are numerous methods, with Youden's *J*-statistic and the point closest to the top-left part, i.e. (0, 1), of the ROC curve plot being two frequently used approaches [97].
- The AUC can be insensitive to detection of model improvement. Two widely accepted measures that overcome this issue are the *net reclassification improvement (NRI) and integrated discrimination improvement (IRI)* [98].

DATA REPORTING

- *Summary statistics* should be reported on subjects (e.g. age, gender, procedures performed and comorbidities), making use of tables where possible.
- *Repetition of summary data* should be avoided by not presenting the same data in multiple formats. For example, data reported in a table can be referred to in the main text without duplication of summary statistics.
- If *different interventions* are being considered, then characteristics should be summarized within each different arm along with an appropriate statistical comparison.
- Averages should always be presented with an appropriate *measure of variability* (e.g. SD or the 25th to 75th percentile interval).
- The term *'average'* should be avoided when used in reference to actual numerical values. Instead, report the actual statistic being reported or compared. For example, when stating 'the average was 55%', it is not clear whether this is a mean, median or otherwise, which might have implications should the distribution be skewed.

- If not implicit, then authors should explicitly *state what summary statistics are being reported*. For example, reporting that 'age in a sample was 65 years (43, 78)' is unhelpful as it is not clear (i) whether the average age was a mean, median or otherwise; (ii) whether the interval denotes the minimum and maximum, the first and third quartile, or otherwise. Extra care is especially required for studies reporting SDs and SEs.
- When reporting summary statistics, it should be clear how these were *calculated if missing data were present*. For example, whether they are case-complete statistics, or maybe calculated following an imputation.

Continuous data

- *Continuous variables* (e.g. age, BMI) can be presented as mean and SD, or if there is evidence of the distribution of data being skewed (can be gauged using a histogram for example), then a useful summary is the median with the 25th to 75th percentile interval (equivalent to the 1st to 3rd quartile interval). In most situations, length-of-stay data will be positively skewed. Note that technically the interquartile range (IQR) is defined as the 3rd quartile minus the 1st quartile (i.e. a single number); however, the interval is the preferred statistic for reporting as it conveys more information.
- *Descriptive statistics should not be reported as inferential statistics*. For example, the mean patient age should be reported with a SD, not a SE or CI.
- While the *minimum and maximum* values are informative, they are easily influenced by outlier values and should therefore not be used in place of the 25th to 75th percentile interval when describing non-normal distributions, but rather in addition to it.
- When reporting *associations between two continuous variables* (e.g. Pearson's sample correlation; simple linear regression), a scatterplot of data should be presented, possibly including a fitted regression line/curve. A correlation coefficient alone can mask complex relationships (see Fig. 1).
- When *defining thresholds*, ensure that all values are captured. For example, if defining two groups as <50% ejection fraction and >50% ejection fraction, one of them needs to include an equality so that an ejection fraction of exactly 50% is captured, e.g. ≤50%.

Categorical data

- *Categorical and binary variables* should be reported as counts and percentages. Reporting counts without percentages makes comparison between groups, e.g. different treatment arms, with unequal group sizes (denominators) difficult.
- The *denominator for percentages* should always be made clear if not implicit. For example, if reporting 70% of 200 patients had valve replacement surgery, and 50% had a mechanical valve prosthesis, then it should be clear whether this is 50% of all patients (i.e. denominator = 200) or 50% of patients undergoing valve surgery (denominator = 140).
- *Ordinal variables* such as New York Heart Association dyspnoea grade should be reported as counts and percentages, not as mean and SD.
- Avoid mixing *percentages and proportions* when reporting data. Additionally, avoid using the terms interchangeably to prevent misinterpretation.
- *Always check that numbers add up correctly*. This is one of the most frequently encountered errors found in tables within manuscripts submitted. For example, if there are 50 patients, and it is reported that 30 had a mitral valve repair, and 18 had a mitral valve replacement, then what procedures did the other 2 patients have?
- *Always check that percentages have been calculated correctly*. Rounding errors are a common error. For example, 1/11 = 9.1%, not 9.0% when rounded to 1 decimal place.

Presentation and notation

- Use of the '±' *symbol for reporting SD and SE values* is confusing unless explicitly stated [100]. Therefore, it is recommended that authors do not use it and instead use parenthesis with a statement on what precision measurement is being used; for example, replace 5.6 ± 0.5 with 5.6 (SD: 0.5) or 5.6 (SE: 0.5) as appropriate.
- Use of the *dash symbol for reporting intervals* (including CIs, minimum and maximum ranges, and percentile intervals) can be confusing, especially in the presence of negative data values. Therefore, it is recommended authors use commas or the word 'to'; for example replace (95% CI: −3.4 – −1.5) with (95% CI: −3.4, −1.5) or (95% CI: −3.4 to −1.5).
- *EJCTS* convention is to use points for *decimal marks*, not commas. This applies to all numerical reporting in the text, tables and figures.
- The *rounding* of summary statistics for patient data such as means and SDs should be guided by the precision of the underlying data. Generally, for example, age, BMI and left ventricular ejection fraction, do not need to be reported to more than 1 decimal place. Biomarker data, for example, might be recorded to a higher precision, and any rounding of summary statistics should reflect this. Percentages can generally be reported to 0 or 1 decimal place without misrepresentation. When the denominator is <100, percentages should always be rounded to 0 decimal places (i.e. as an integer).
- *Units of measurement* should be reported throughout the manuscript. For example, serum creatinine might be measured in milligrams per deciliter or micromoles per litre.
- Avoid using the *tilde symbol (∼)* as shorthand for the word 'approximately'.

Figures and tables

- In general, when reporting many items of data, *tables and figures* are the preferred presentation format rather than text. Figures are in particular preferred for very large amounts of data. When there are very few data, e.g. 3 or 4 mean values, then figures are a waste of space, unless additional information is to be overlaid (e.g. the raw values).
- Provide *footnotes for tables and figures* to define any abbreviations, acronyms or symbols used.
- *Flowcharts*—a minimum requirement for randomized controlled trials and meta-analyses—are a useful construct for observational data studies with inclusion and exclusion criteria, and those with subgroup analyses.
- *Bar charts are preferred over pie charts* for reporting data on proportions with multiple groups.

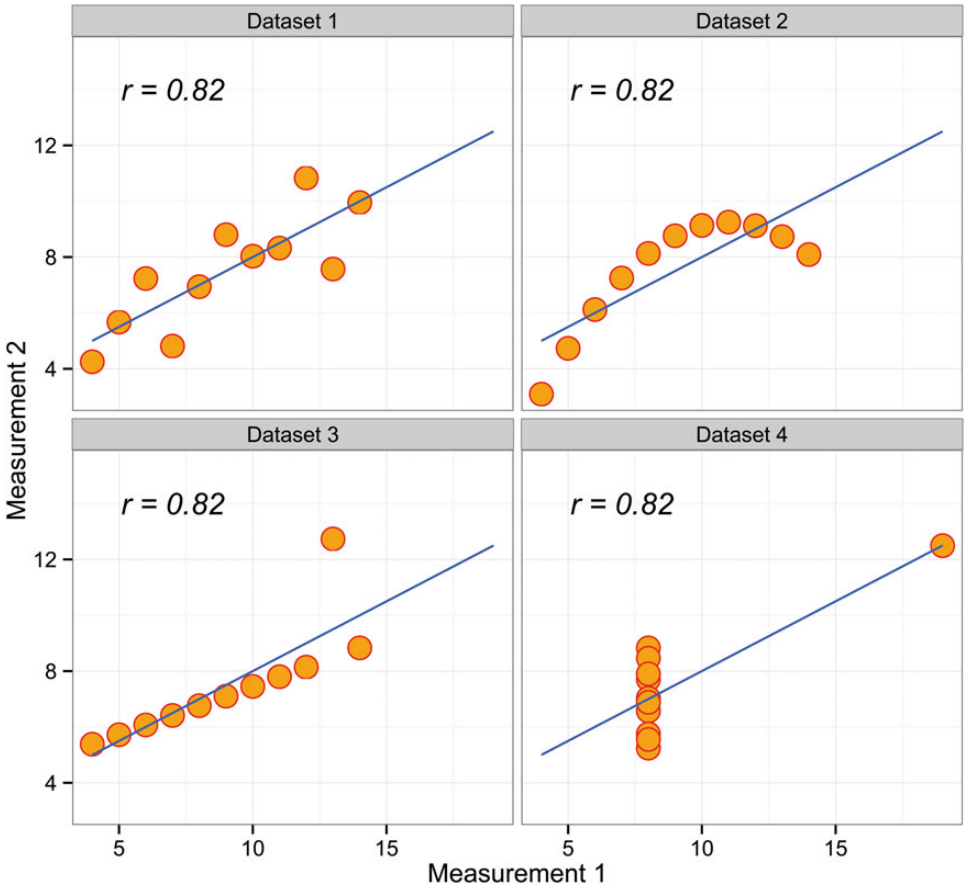

Figure 1: Scatterplots of four different datasets known as Anscombe's quartet [99]. Each dataset consists of 11 data points (orange points) and have nearly identical statistical properties, including means, sample variances, Pearson's sample correlation (denoted as *r* in the figure), and linear regression line (blue lines). This well-known quartet highlights the importance of graphing data prior to analysis, and why statistical reviewers often ask for such graphs to be made available.

RESULTS SECTION

- *Selective reporting* (or filtering of results) is to be avoided. If a study was designed to test a fixed number of hypotheses, then all test results should be reported regardless of statistical significance.
- *Deviations from study design* should be reported. This includes 'going off-piste' and performing additional hypothesis tests (for example, subgroup analyses, which are likely to include far fewer subjects; or combining endpoints such as different post-operative complications in a single composite indicator) on the basis of primary findings.
- Appropriate *effect sizes* should be reported. A *P*-value is not an effect size. Standard effect sizes for studies in the *EJCTS* and *ICVTS* include:
  - *Odds ratios (ORs)* for binary data and logistic regression (with a logit-link function).
  - Relative risks (or risk ratios) for binary data and logistic regression (with a log-link function).
  - *Hazard ratios (HRs)* for Cox proportional hazards regression.
  - *Regression coefficients* for risk prediction models or multiple linear regression.
  - *Differences in means* for continuous measurements and proportions.
  - *Differences in survival* at a fixed time point.
- A careful distinction should be made between reporting *relative and absolute effect sizes*. For example, reporting a HR vs difference in median survival, or an OR vs difference in proportions.
- *It is desirable to report the size of the difference in a clinically meaningful way* in addition to *P*-values instead of just the latter for primary hypothesis tests. For example, one might report incidence differences of '16 vs 40%; *P* = 0.008' rather than just '*P* = 0.008'.
- Explicitly *state what the effect size is*, as opposed to writing an acronym. For example, state 'odds ratio (OR)' at first instance in the Abstract and the main text rather than just 'OR'. It is useful to note that two frequently observed errors are authors: (i) interchanging 'OR' and 'HR' (i.e. OR and HRs); and (ii) confusing both 'OR' and 'HR' with 'RR' (relative risk).
- *Regression models should, in most circumstances, be reported in full*. That is, all coefficients (or corresponding effect sizes) should be summarized.
- If appropriate, regression model *intercepts* should be reported. For example, these are required if the regression model is used as a risk prediction model or to estimate a lung topology outcome from spirometry data.
- *CIs are preferred* over *P*-values. For example, reporting that a particular treatment was statistically significantly associated with an increase in length of hospital stay is less informative than reporting that the mean increase in length of hospital stay was 1.3 days (95% CI: 0.2–2.4).
- CIs convey the precision of an estimate. A *wide CI*, whether statistically significant or not, should be interpreted with caution.

- For most studies it is sufficient to *report P-values* to 3 decimal places if <0.10, and 2 decimal places if ≥0.10. If a *P*-value is <0.001, then reporting '*P* < 0.001' is sufficient unless there are an unusually large number of hypothesis tests (e.g. genetic differential expression data). Avoid reporting unnecessarily high degrees of precision or only reporting *P*-values to 1 decimal place.
- When *reporting P-values*, avoid the following:
  - *Arbitrary bounds*, for example *P* < 0.05; *P* > 0.4 or '*P* less than 0.05'.
  - Reporting *'n.s.' or 'NS'* (i.e. 'non-significant') merely because it is greater than some arbitrary value.
  - *P = 0.000* when rounded.

  Instead report the exact *P*-value as described earlier.
- The most appropriate *precision of effect sizes* will vary by study. ORs and HRs can often be reported adequately to 1–2 decimal places; however, clinical risk prediction models might want to consider reporting coefficients to a greater precision.
- When a study involves following patients over time, the *follow-up data should be succinctly summarized*. This might include the median (and range) follow-up time and/or total person-years. Additionally, the date of the last follow-up check and whether patients were lost during the follow-up (and why) should be reported. Note that if a median follow-up time is provided, the authors should specify how they calculated this. Just taking the median time to event or censoring in a group is often not sufficient as this is biased by treatment group differences. It is preferable to use an inverse Kaplan–Meier by having censoring time as an event and censor patients with an event at the time of the event.

Figures (general)

- Figures should be provided in a *high-resolution* format (see *EJCTS* and *ICVTS* 'Instruction to Authors': http://www.oxfordjournals.org/our_journals/ejcts/for_authors/manuscript_instructions.html) to enable clarity in presentation. Consideration should be given to the font size, point size and line widths.
- Authors should consider making use of *colour figures* where appropriate. Note that the printing of colour figures is free of charge in the *EJCTS* and *ICVTS*. However, please note the Editor may use his discretion when deciding which figures to publish in colour.
- All *axes* of figures should be clearly labelled, including specifying any units of measurement.
- If *P-values* or other statistical results are reported in a figure, a description of the hypothesis and methodology should be provided in the figure legend.
- If *asterisks* (or other symbols) are used in a plot, for example to indicate a particular significance threshold, they should be defined in the legend, and preferably consistent throughout the manuscript.
- *Error bars* are sometimes shown in plots. If 'error bars' are to be added, it must be clearly stated in the figure legend what they denote, e.g. ±1 SD, ±1 SE, 95% CI or otherwise.
- If figures are presented on *transformed scales* (e.g. log paper) or axes begin from an arbitrary origin (e.g. having a Kaplan–Meier survival curve only shown on 50–100% cumulative survival), then this should be drawn to the reader's attention in the figure legend.
- *'Dynamite plots'* (bar charts with error bar lines) should be avoided [101]. They hide important data and are not always suitable. Dot plots and box-and-whiskers plots are useful alternatives when sample sizes are small and large, respectively.
- *Three-dimensional bar charts and pie charts and other 'special effects' charts*, all of which are ubiquitous to desktop office software suites, should not be used in any circumstance, as they are difficult to read and contrast.

Figures (Kaplan–Meier curves; Fig. 2)

- When there is potential for ambiguity, it is preferred that the *horizontal axis label* of a Kaplan–Meier graph should read 'Time from ... (units)', where the '...' denotes the time origin (e.g. randomization, time of surgery, discharge) and 'units' denotes the units of the axis (e.g. months, years).
- The *horizontal axis should be naturally discretized* according to the length of follow-up. For example, it would be most appropriate to use units of years rather than blocks of 100 days in a Kaplan–Meier curve extending out to 5 years.
- *Displaying 'error bars' at fixed time points* on Kaplan–Meier curves is not particularly useful, especially when there are multiple groups (strata) as they overlap and add unnecessary distraction. If a measure of precision is required, CI bands should be included.
- Kaplan–Meier graphs should be displayed with a table showing the *number of patients at risk* at a sequence of regularly spaced time points. The table should be aligned with the horizontal axis (Fig. 2).
- Censoring information should be provided using *tick marks* (or similar). These should also be stated in the graph, for example using a legend or noted in the figure description. In some situations, e.g. analysis of large national registries, tick marks might be omitted.
- There is a *greater uncertainty about the survival towards the end* of the Kaplan–Meier curve. To avoid misinterpretation about this region of the graph, the horizontal [time] axis might be truncated when there remain only a few patients at risk.

DISCUSSION SECTION

- *The discussion and conclusions should be transparent* if a finding was based on pre-specified analysis or an exploratory *post hoc* finding. In the case of the latter, it should generally be viewed as exploratory and of a hypothesis-generating nature requiring confirmation by a further study.
- A statistically significant effect demonstrates an association, but does not constitute evidence of *causation*. In general, causation is difficult to infer, except perhaps in randomized controlled trials.
- Effects found to be statistically significant should be evaluated for *clinical significance*. For example, a study that finds surgery reduces the mean time patients spend in intensive care by 0.1 days might be statistically significant, but might not have clinical significance.
- *P-values only provide evidence against a hypothesis*, never evidence in favour of it. In other words absence of evidence is never evidence of absence. For example, a *P*-value of 0.60 only tells us that there is insufficient evidence for an effect, which might be due to either no effect being present, or insufficient information in the data sample due to a small sample size, large variability or both [103].
- *A P-value does not equal the probability that a null hypothesis is true*. Therefore, *P*-values cannot be used to rank evidence. For example, if a Hosmer–Lemeshow test was applied to evaluate two separate risk prediction models, then the size of the *P*-values does not confer which is the 'better' model.

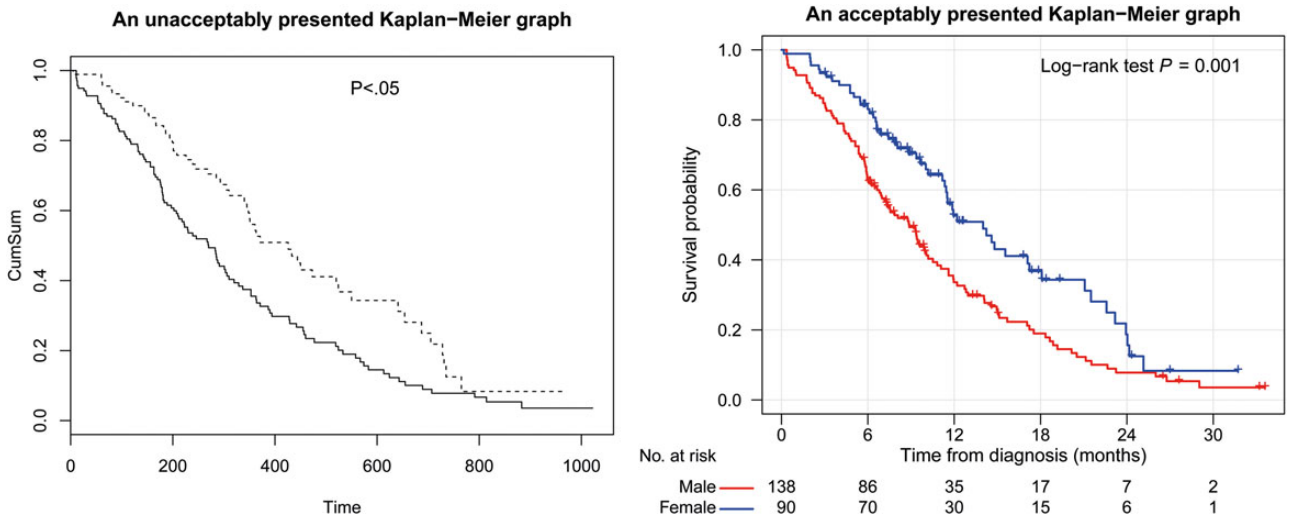

Figure 2: Kaplan–Meier graphs showing the survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. The left panel shows an unacceptably presented graph. Note how it lacks: (i) a key stating what groups the lines correspond to (although this information might be conveyed in the figure legend); (ii) units for the horizontal axis; (iii) reference to the time origin of the study (which is not always implicit from the study); (iv) a table reporting the number of patients at risk for a discrete set of time points; (v) a clearly labelled vertical axis; (vi) marks denoting censoring times; (vii) an exact *P*-value, which is only reported as being below a particular threshold; (viii) a natural subdivision of the horizontal axis, which is broken up into arbitrary time units of 200 (days). The right panel shows an acceptably presented graph. In addition to overcoming the shortcomings of the left panel, it also makes use of colour lines for differentiating between the curves. These data were accessed in the R survival package (version 2.37–7) [102].

- *Study weaknesses* should be well thought out. Authors should go beyond merely commenting on the study sample size and whether it was observational data or not. Consideration should be given towards the potential for study inclusion/exclusion criteria to distort the results; methodology applied; any assumptions made (e.g. missing data); the sample size in contrast to the number of events; any unmeasured confounders, especially those that might be strong; violation of statistical assumptions etc. The implications of study weaknesses on the findings should be determined.
- An attempt should be made to explain *unexpected results* or those that contradict existing findings with reference to the data and analysis design.
- Conclusions should only be based on findings in the study that can be backed up by study data and analyses. General statements beyond study findings should not be made.

ACKNOWLEDGEMENTS

The authors thank Peter Diggle (Institute of Infection and Global Health, University of Liverpool) for his input during the preparation of these guidelines and also for reviewing an earlier draft.

Funding

GLH acknowledges support by the Medical Research Council (MRC) Health e-Research Centre, Farr Institute of Health Informatics (Grant: MR/K006665/1).

Conflict of interest: none declared.

REFERENCES

[1] Dunning J. How to complete a review for the European Journal of Cardio-Thoracic Surgery and the journal Interactive CardioVascular and Thoracic Surgery. Eur J Cardiothorac Surg 2012;41:242–7.
[2] Guidelines for data reporting and nomenclature for The Annals of Thoracic Surgery. Ann Thorac Surg 1988;46:260–1.
[3] Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. Br Med J 1983;286:1489–93.
[4] Bland JM. An Introduction to Medical Statistics. 3rd edn. Oxford, UK: Oxford University Press, 2000.
[5] Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991.
[6] Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D *et al.* Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000;283:2008–12.
[7] Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D *et al.* Consolidated Health Economic Evaluation Reporting Standards (CHEERS)—explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Heal 2013;16:231–50.
[8] Altman DG, Bland JM. Statistics notes: How to randomise. Br Med J 1999; 319:703–4.
[9] Day SJ, Altman DG. Statistics notes: Blinding in clinical trials and other studies. Br Med J 2000;321:504.
[10] Kang M, Ragan BG, Park J-H. Issues in outcomes research: an overview of randomization techniques for clinical trials. J Athl Train 2008;43:215–21.
[11] Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: Chance, not choice. Lancet 2002;359:515–9.
[12] Dafni U. Landmark analysis at the 25-year landmark point. Circ Cardiovasc Qual Outcomes 2011;4:363–71.
[13] Akins CW, Miller DC, Turina MI, Kouchoukos NT, Blackstone EH, Grunkemeier GL *et al.* Guidelines for reporting mortality and morbidity after cardiac valve interventions. Eur J Cardiothorac Surg 2008;33: 523–8.
[14] Kappetein AP, Head SJ, Généreux P, Piazza N, van Mieghem NM, Blackstone EH *et al.* Updated standardized endpoint definitions for transcatheter aortic valve implantation: the Valve Academic Research Consortium-2 consensus document (VARC-2). Eur J Cardiothorac Surg 2012;42:S45–60.
[15] Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 2003;289:2554–9.
[16] Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. Br Med J 2010;341:c3920.
[17] Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16:9–13.

[18] Roques F. The logistic EuroSCORE. Eur Heart J 2003;24:1–2.
[19] Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR *et al*. EuroSCORE II. Eur J Cardiothorac Surg 2012;41:1–12.
[20] Laine C, de Angelis C, Delamothe T, Drazen JM, Frizelle FA, Haug C *et al*. Clinical trial registration: looking back and moving ahead. Ann Intern Med 2007;147:275–7.
[21] Knol MJ, Groenwold RHH, Grobbee DE. P-values in baseline tables of randomised controlled trials are inappropriate but still common in high impact journals. Eur J Prev Cardiol 2012;19:231–2.
[22] Peters TJ. Multifarious terminology: multivariable or multivariate? Univariable or univariate? Paediatr Perinat Epidemiol 2008;22:506.
[23] Wormuth DW. Actuarial and Kaplan-Meier survival analysis: there is a difference. J Thorac Cardiovasc Surg 1999;118:973–5.
[24] Altman DG, Bland JM. Quartiles, quintiles, centiles, and other quantiles. Br Med J 1994;309:996.
[25] Sullivan LM. Repeated measures. Circulation 2008;117:1238–43.
[26] Diggle PJ, Heagerty P, Liang K-Y, Zeger S. Analysis of Longitudinal Data. 2nd edn. Oxford, UK: Oxford University Press, 2013.
[27] Gelman A, Hill J. Data Analysis Using Regression and Multilevel/ Hierarchical Models. 1st edn. New York: Cambridge University Press, 2007.
[28] Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol 2001;1:6.
[29] Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. Br Med J 2001;323:1123–4.
[30] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. Br Med J 1995;310:170.
[31] Hart A. Mann-Whitney test is not just a test of medians: differences in spread can be important. Br Med J 2001;323:391–3.
[32] Campbell I. Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. Stat Med 2007;26:3661–75.
[33] Van der Heijden GJMG, Donders RRT, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006;59:1102–9.
[34] Knol MJ, Janssen KJM, Donders RRT, Egberts ACG, Heerdink ER, Grobbee DE *et al*. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 2010;63:728–36.
[35] Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8: 3–15.
[36] Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991;10:585–98.
[37] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Br Med J 2009;338:b2393.
[38] Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;79:340–9.
[39] Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer, 2001.
[40] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935–42.
[41] Flom PL, Cassell DL. Stopping stepwise: why stepwise and similar selection methods are bad, and what you should use. In: Mitchell R, Williams C (eds). Northeast SAS Users Gr. Annu. Conf., Baltimore, 2007.
[42] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41.
[43] Draper NR, Smith H. Applied Regression Analysis. 3rd edn. New York: Wiley, 1998.
[44] Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. Biometrika 1990;77:147–60.
[45] Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 1999; 28:964–74.
[46] Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd edn. New Jersey: John Wiley & Sons, Inc., 2000.
[47] Slinker BK, Glantz SA. Multiple linear regression: accounting for multiple simultaneous determinants of a continuous dependent variable. Circulation 2008;117:1732–7.
[48] Altman DG, Matthews JN. Statistics notes: interaction 1: heterogeneity of effects. Br Med J 1996;313:486.
[49] Gelman A, Hill J. Data Analysis Using Regression and Multilevel/ Hierarchical Models. Cambridge: Cambridge University Press, 2006.
[50] Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med 1998;17:1623–34.
[51] Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics 1983;39:499–503.
[52] Steyerberg EW, Schemper M, Harrell FE Jr. Logistic regression modeling and the number of events per variable: selection bias dominates. J Clin Epidemiol 2011;64:1464–5.
[53] Vittinghoff E, Mcculloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007;165:710–8.
[54] Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982; 115:92–106.
[55] Collett D. Modelling Survival Data in Medical Research. 2nd edn. Boca Raton, FL: Chapman & Hall/CRC, 2003.
[56] Grambsch PM, Therneau TM. Proportional hazard tests and diagnostics based on weighted residuals. Biometrika 1994;81:515–26.
[57] Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. 2nd edn. New York: Springer, 2003.
[58] Oakes D, Peterson DR. Survival methods: additional topics. Circulation 2008;117:2949–55.
[59] Ahmed FE, Vos PW, Holbert D. Modeling survival in colon cancer: a methodological review. Mol Cancer 2007;6:15.
[60] Crowley J, Hu M. Covariance analysis of heart transplant survival data. J Am Stat Assoc 1977;72:27–36.
[61] Kleinbaum DG, Klein M. Survival Analysis. 3rd edn. New York: Springer, 2012.
[62] Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. Br J Cancer 2004; 91:1229–35.
[63] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequancy, and measuring and reducing errors. Stat Med 1996;15:361–87.
[64] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56:337–44.
[65] Finkelstein DM. Comparing survival of a sample to that of a standard population. J Natl Cancer Inst 2003;95:1434–9.
[66] Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer, 2009.
[67] Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23: 2567–86.
[68] Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y *et al*. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.
[69] Sullivan LM, Massaro JM, D'Agostino Snr RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med 2004;23:1631–60.
[70] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. Br Med J 2009;338: 1432–5.
[71] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG *et al*. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691–8.
[72] Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Crit Care Med 2007;35:2052–6.
[73] Marcin JP, Romano PS. Size matters to a model's fit. Crit Care Med 2007; 35:2212–3.
[74] Paul P, Pennell L, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. Stat Med 2013;32: 67–80.
[75] Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014;33:517–35.
[76] Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562–5.
[77] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.
[78] Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58:475–83.

[79] Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. J Thorac Cardiovasc Surg 2007;134:1128–35.
[80] Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2006;15:199–236.
[81] Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083–107.
[82] Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiol Drug Saf 2008;17: 1202–17.
[83] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46:399–424.
[84] Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med 2006;25:2084–106.
[85] D'Agostino RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81.
[86] Stuart EA. Developing practical recommendations for the use of propensity scores: discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. Stat Med 2008;27:2062–5; discussion 2066–9.
[87] Hill J. Discussion of research using propensity-score matching: comments on ' A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003 ' by Peter Austin, Statistics in Medicine. Stat Med 2008;27:2055–61.
[88] Imbens GW. The role of the propensity score in estimating dose-response functions. Biometrika 2000;87:706–10.
[89] Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. Epidemiology 2013;24:401–9.
[90] D'Agostino RB Jr. Propensity scores in cardiovascular research. Circulation 2007;115:2340–3.
[91] Normand S-LT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD *et al*. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. J Clin Epidemiol 2001;54:387–98.
[92] Bland JM, Altman DG. Matching. Br Med J 1994;309:1128.
[93] Altman DG, Bland JM. Statistics notes: diagnostic tests 1: sensitivity and specificity. Br Med J 1994;308:1552.
[94] Altman DG, Bland JM. Statistics notes: diagnostic tests 2: predictive values. Br Med J 1994;309:102–102.
[95] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
[96] Delong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
[97] Perkins NJ, Schisterman EF. The inconsistency of 'optimal' cut-points using two ROC based criteria. Am J Epidemiol 2006;163:670–5.
[98] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157–72.
[99] Anscombe FJ. Graphs in statistical analysis. Am Stat 1973;27:17–21.
[100] Jaykaran . 'Mean±SEM' or 'Mean (SD)'? Indian J Pharmacol 2010;42:329.
[101] Tobias A. Dynamite plunger plots should not be used. Occup Environ Med 1998;55:361–2.
[102] Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New Jersey: Springer, 2000.
[103] Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. Br Med J 1995;311:485–485.